Probability and Intro Statistics I

Nick Litombe

Harvard & www.alliance4ai.org

November 23, 2019

Events, Outcomes, Sample spaces

An outcome is a possible result of an experiment. For example, tossing a single coin has two outcomes $\{H\}, \{T\}$.

The set of all possible outcomes is called a sample space.

An event is a set of outcomes of an **experiment** which means it's a subset of a **sample space**. An event **occurred** if an actual outcome of an experiment is in A.

An empty set is the set of no outcomes $\varnothing$.

Examples

Experiment 1: Toss a coin once

Outcomes: HEADS or TAILS

Some events:

- Set of outcomes of heads H
- Outcomes of tails T
- Outcomes of heads or tails H, T
- Outcomes of neither heads nor tails $\varnothing$

Examples

Experiment 2: Toss a coin twice OR toss two coins at the same time

Outcomes: HH, HT, TH, TT

Some events:

- Set of outcomes with first coin/toss heads
- Set of outcomes with all tails
- Set of outcomes with all heads
- Set of outcomes with no tails

Examples

Experiment 3: Throwing a 6-sided die

Outcomes: $\{1, 2, 3, 4, 5, 6\}$

Some events:

- Set of outcomes of even numbers
- Outcomes of odd numbers
- Outcomes that sum to 7
- Outcomes that belong to $1, 5$

Examples

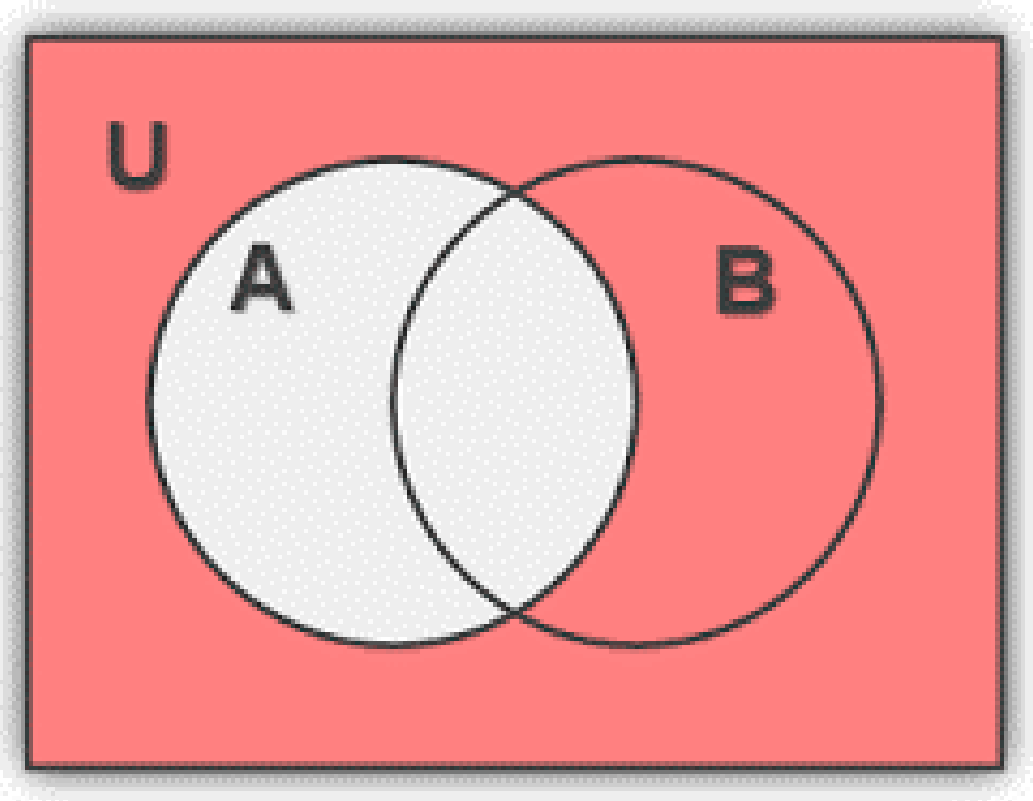
Experiment 4: Select a card from a 52-card deck

Outcomes: $\{A, K, Q, J, 10, 9, 8, 7, 6, 5, 4, 3, 2\} \times \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$

Some events:

- all aces
- all red suits
- all numbered cards
- all cards with values greater than or equal to ten

Probability Sum



Consider two events, A and B which are two sets of possible outcomes, then,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1}$$

Types of Probability

Consider two events, A and B which are two sets of possible outcomes, then,

Joint probability $P(A, B) = P(A \cap B)$

- Probability of events A and B which requires outcomes common to both A and B to be observed.

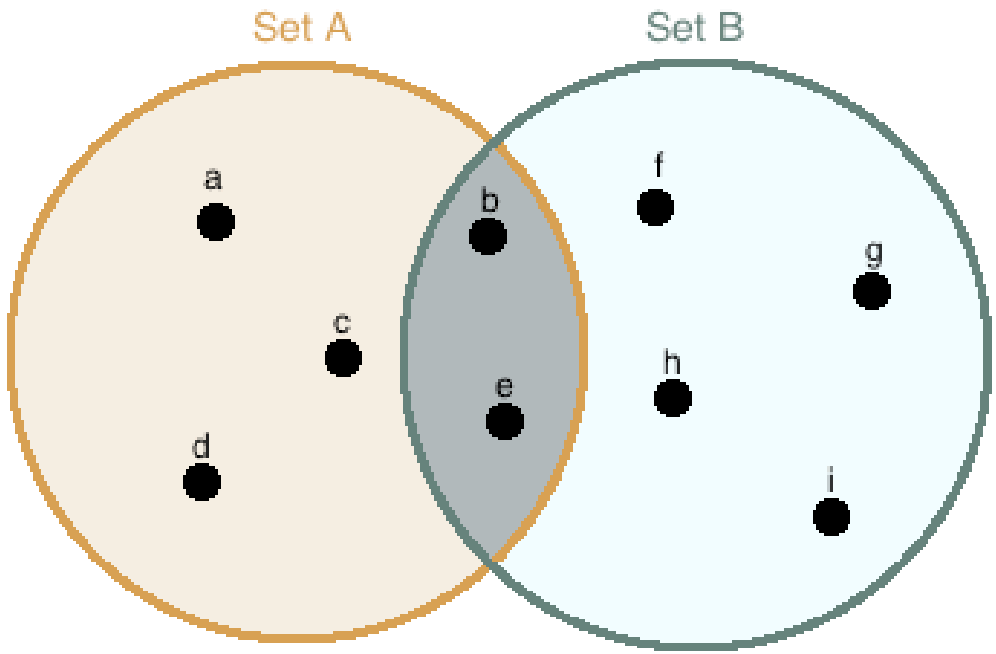
Conditional probability $P(A \mid B) = \frac{P(A,B)}{P(B)}$

- Probability measures normalized in new sample space B.
- $P(A \mid B) \neq P(B \mid A)$

Marginal probability $P(A)$

- Averaging over outcomes from event B

Conditioning



Conditional Probability & Bayes' Rule

Conditional probability $P(A \mid B) = \frac{P(A,B)}{P(B)}$

$P(A, B) = P(A \mid B) \cdot P(B)$

$P(B, A) = P(B \mid A) \cdot P(A)$

Bayes' Rule

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Independence

Consider two events, A and B. What does it mean to say these two events are independent?

We can answer this question because we now know all about conditional probabilities.

$P(A, B) = P(B)P(A)$

$P(A \mid B) = P(A)$

So B provides no information in calculating probability of A

Equivalently,

$P(B \mid A) = P(B)$

Odds and Ends

Let's make the marginal probability more explicit with what's known as the Law of Total Probability.

$P(A) = \sum_n P(A \mid B_n)P(B_n)$

For definition purposes, note that:

$odds(A) = \frac{P(A)}{P(A^c)}$

In Logistic Regression, we fit the Log-Odds ratio to a linear relation.

Random Variables

Random variables are primary objects in Probability and Statistics so let's offer a (stripped down) definition:

A random variable is a function $X : \Omega \to E$ from a set of possible outcomes Ω to a space E. E will be assumed to be R, the real numbers.

Examples

A coin toss:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \text{heads}, \\ 0, & \text{if } \omega = \text{tails}. \end{cases} \tag{2}$$

$$f_X(x) = P(X = x) = \begin{cases} \frac{1}{2}, & \text{if } x = 1, \\ \frac{1}{2}, & \text{if } x = 0, \end{cases} \tag{3}$$

When you toss a coin, you don't know for sure if it will be heads or tails, only that it is one of these possible outcomes. So the outcome of the experiment is a Random Variable.

Examples of Random Variables

Experiment: Select a 6 cards from a 52-card deck

single outcomes:
$\{A, K, Q, J, 10, 9, 8, 7, 6, 5, 4, 3, 2\} \times \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$

Some Random Variables (r.v.'s):

- the number of spades suits in your 6-card selection
- the sum of the values of the hand you selected
- the difference between the number of red vs. black cards
- the maximum value of the cards selected

Note: A function of an r.v is also an r.v.

Expectation

$$\mathrm{E}[X] = \sum_{x} x \cdot P(X = x) \tag{4}$$

- We can define an average value that a random variable takes
  - Mean, Expected Value or expectation all the same thing
- Expected value of a 6-sided die
- Expected value of a coin toss

Linearity of Expectations and LOTUS

The Linearity of expectation of random variables holds for any r.v.'s. X, Y which are dependent or independent.

$$\begin{aligned} \mathrm{E}[X+Y] &= \mathrm{E}[X] + \mathrm{E}[Y] \\ \mathrm{E}[cX] &= c\,\mathrm{E}[X] \\ \mathrm{E}[g(X)] &= \sum_x g(x) P(X=x) \end{aligned}$$

- This finds applications in calculating the distribution of a sum of random variables which occurs often in statistics
- Also allows to find expectation of more complicated distributions if they are the sum of several random variables from distributions for which we know the mean

Variance

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \tag{5}$$

The spread or degree of uncertainty in the distribution of X. It's the mean squared deviation. The standard deviation σ_x is the square root of the variance.

- Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \\
&= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}(X) + [\mathrm{E}(X)]^2\right] \\
&= \mathrm{E}(X^2) - \mathrm{E}[2X\,\mathrm{E}(X)] + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\,\mathrm{E}(X) + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - 2[\mathrm{E}(X)]^2 + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2
\end{aligned}$$

Covariance

$$\operatorname{Cov}(X_i, X_j) = \operatorname{E}(X_i X_j) - \operatorname{E}(X_i)\operatorname{E}(X_j) \tag{6}$$

Covariance is a measure of the joint variability of two variables.

- Proof:

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right] \\
&= \operatorname{E}\left[XY - X\operatorname{E}[Y] - \operatorname{E}[X]Y + \operatorname{E}[X]\operatorname{E}[Y]\right] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] - \operatorname{E}[X]\operatorname{E}[Y] + \operatorname{E}[X]\operatorname{E}[Y] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y],
\end{aligned}$$

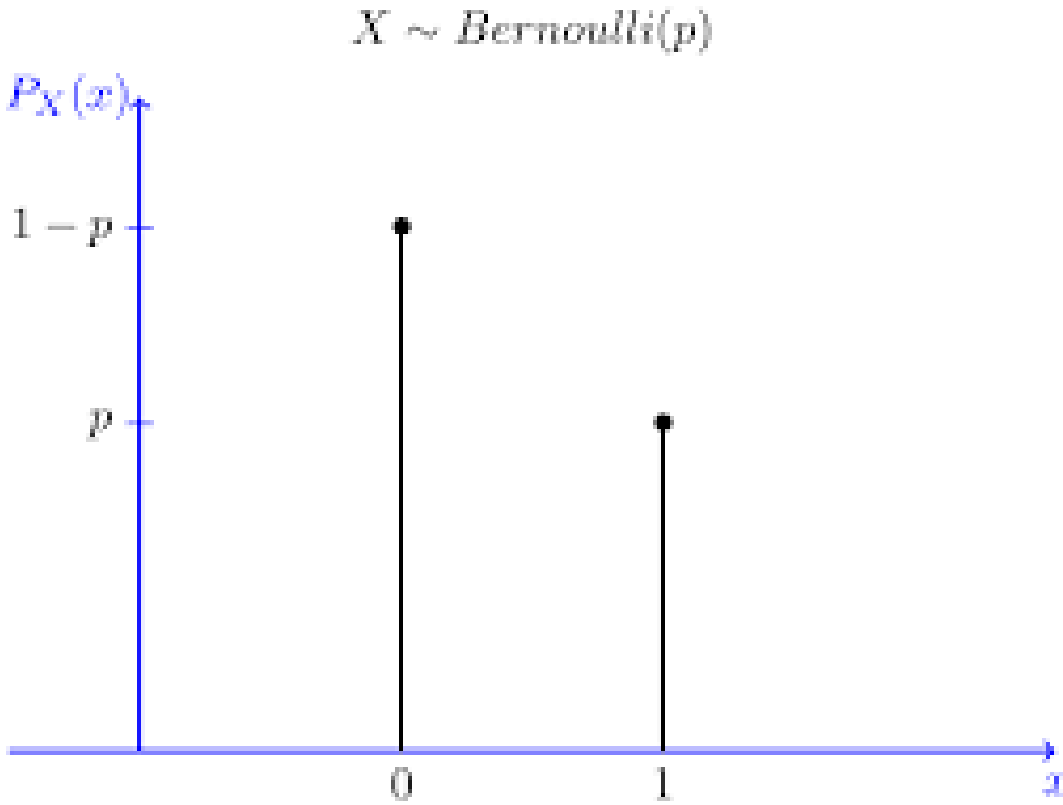
Correlation

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{7}$$

Pearson's correlation measures the linear relationship between random variables and ranges from -1 to 1.

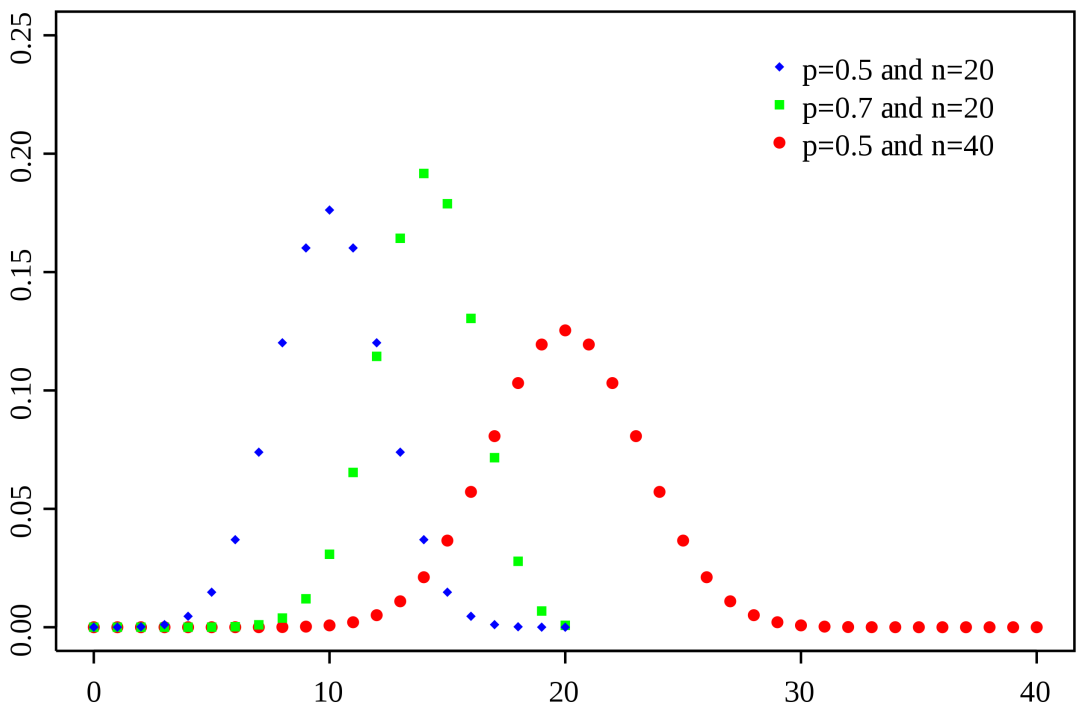
- Just because the correlation between X, Y is zero doesn't mean there's no functional relationship between the two variables.
- Here we are only testing for linearity.

Bernoulli Distributions



- x-axis shows outcome values, and y-axis their probabilities.
- This is for a biased trial or biased coin
- $\mathrm{E}[X] = p$
- $\mathrm{Var}[X] = p(1-p) = pq$

Binomial Distributions



- x-axis is the number of successes or number of heads. The y-axis is the probability.
- If a coin is tossed N times, it can be seen as the sum of N Bernoulli trials or random variables.
- $\mathrm{E}[X] = np$ by linearity of expectation
- $\mathrm{Var}[X] = np(1-p) = npq$

Things to note

- Just because two random variables have the same distribution, does not make them the same
- $X = x$ is an event because it is all those outcomes in sample space that map to x in R space.
- An expected value is a number and a random variable r.v is a function. Don't interchange the two
- Two distributions can have the same mean but be different distributions
- The expectation depends only on the distribution

Variance Relations

Often times we want to know the variance of sums of random variables or scaled random variables.

$$\begin{aligned} \mathrm{Var}[X + c] &= \mathrm{Var}[X] \\ \mathrm{Var}[cX] &= c^2 \, \mathrm{Var}[X] \\ \mathrm{Var}[X + Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Cov}(X, Y) \end{aligned}$$

- Note that if X and Y are independent the variance of their sum is the sum of the variances.
- Var is zero or positive. It's zero if the r.v. is a constant.

i.i.d

- iid == independent, identically distributed
- Each random variable while independent all have the same probability distribution

Statistics

Statistics is the study of data

- We usually have data drawn from a population whose true parameter values we do not know
- We would like to infer or estimate the parameters from a sample
- A sample is a finite amount of data derived from the sample

Probability and connection to Statistics

Statistics is the study of data

- We usually have data drawn from a population whose true parameter values we do not know
- We would like to infer the parameters from a sample
- A sample is a finite amount of data derived from the population
- For example, a coin has a bias, p. This is the population parameter unknown to us
- We flip this coin N times. Our dataset is N-long.
- We've collected our data by a well-defined procedure of flipping a coin.
- How do we use this data sample to infer or estimate what the population parameter is?

Sample mean and Variance of sample mean

Let's take a sample of data: N data points. $X_1, \ldots, X_N$.

Assume **iid** i.e. the same population parameters $\{\mu, \sigma^2\}$. $\bar{X}$ is the sample mean so calculate its expected value.

$$\mathrm{E}[\bar{X}] = \frac{\mathrm{E}[X_1 + \ldots + X_N]}{N} = \frac{N \cdot \mathrm{E}[X_1]}{N} = \mu$$

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}(X_1 + \ldots + X_N) \\ &= N \cdot \mathrm{Var}\left(\frac{X_1}{N}\right) = \frac{\sigma^2}{N} \end{aligned}$$

- We see with more data points decrease the spread or uncertainty in our estimation of the mean.
- We've now broached the subject of the sampling distribution
- $\sigma/\sqrt{N}$ is the standard error

Sample Variance

Let's take a sample of data: N data points. $X_1, \ldots, X_N$.

Assume **iid** i.e. the same population parameters $\{\mu, \sigma^2\}$, but independent of each other, so iid. $\bar{X}$ is the sample mean.

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{i=N} (X_i - \bar{X})^2$$

- It can be shown $\mathrm{E}\left(S^2\right) = \sigma^2$ thereby making it an unbiased estimate of σ^2, the population parameter.

Law of Large Numbers

The Law of Large numbers tells what the mean converges to as $N \to \infty$ under the assumption of iid data.

$$\mathrm{E}[\bar{X}] = \mu$$
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{N}$$

- What's stated above is closer to the Strong Law of Large Numbers
- The weak version is a convergence in probability

Central Limit Theorem

Given i.i.d samples $X_1 \dots X_N$ with mean μ and variance σ^2, for large N:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$
$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$

- The Central Limit Theorem (CLT) is a principal reason why the Normal distribution dominates statistics.
- CLT makes it clear the distribution is a normal distribution

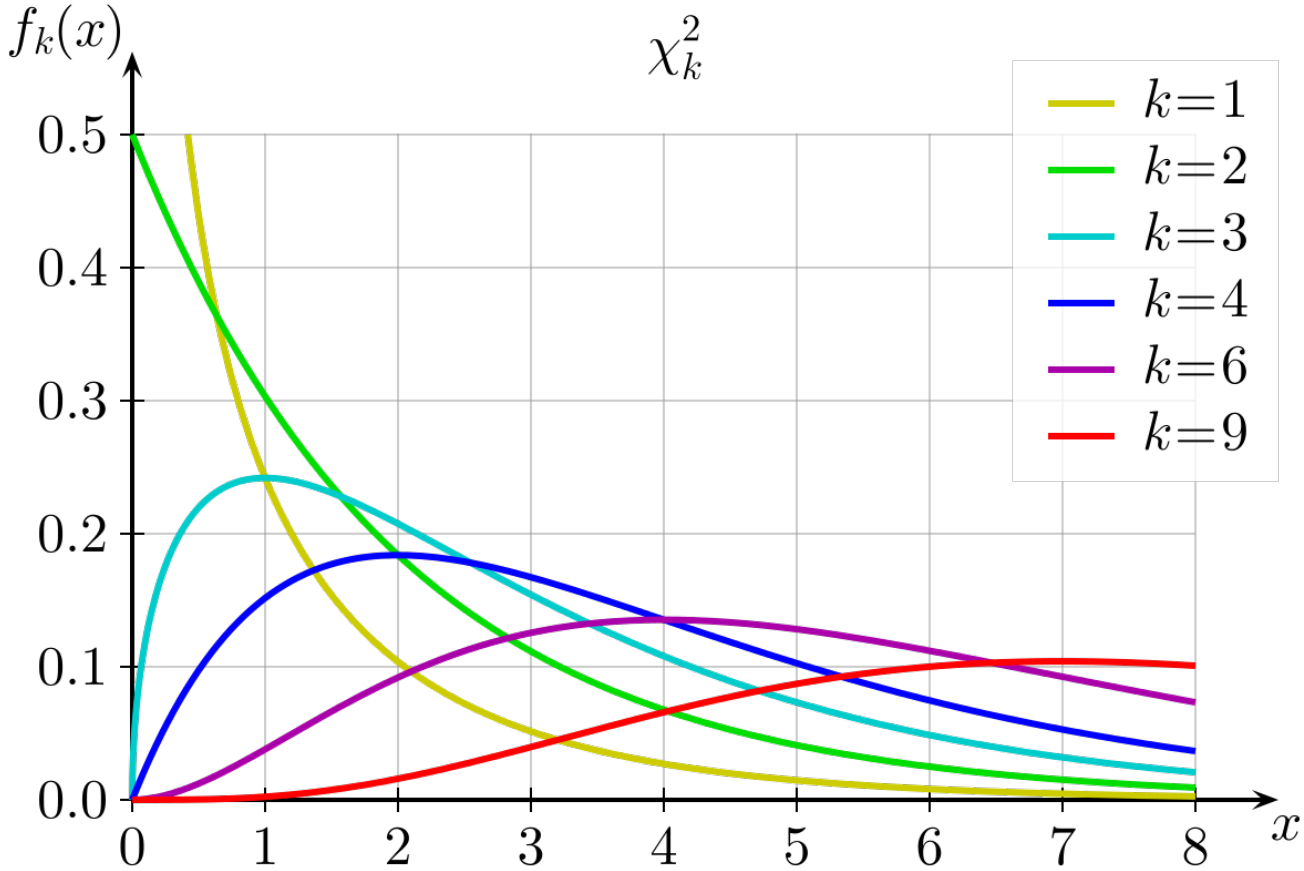
Chi-Squared

Given $Z_1 \ldots Z_N$ i.i.d. are N(0,1), define:

$Z = V_1^2 + \ldots V_N^2$. $Z \sim \chi_N^2$. N is the number of degrees of freedom.

- Z follows a chi-squared distribution with mean N and variance $2N$.
- Chi-squared is important because it is related to the distribution of sample variance
- The sample variance distribution can be used to estimate the true variance of the population.
- $\frac{(N-1)S_N^2}{\sigma^2} \sim \chi_{N-1}^2$. S_N^2 is the sample variance (unbiased).

chi-squared distribution plot



wikipedia.org

t-distribution

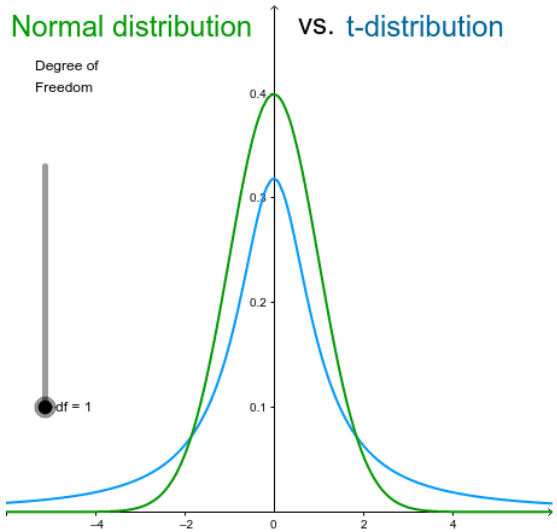
Given $Z \sim N(0,1)$, $V \sim \chi^2_N$ and V is independent of Z then define:

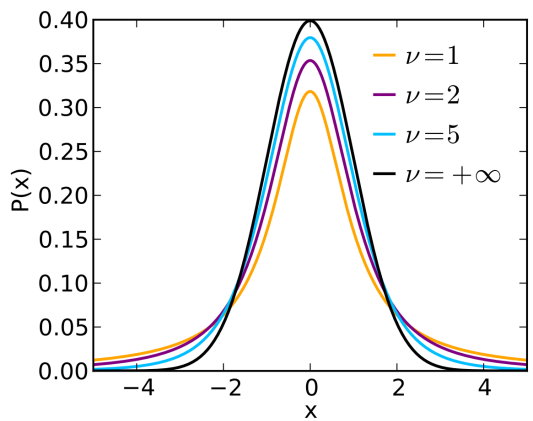
$T = \frac{Z}{\sqrt{V/N}}$.

T is said to be a t-distribution with N degrees of freedom.

- The t-distribution has fatter tails so more higher probability of extreme values
- It converges to the normal distribution as $N \to \infty$
- Because V is a random variable itself a sum of N random variables, if $N = 1$, then the t_1 t-distribution is the Cauchy distribution.

t-distribution plot





- Note that the plot is made for degree of freedom (dof) = 1

geogebra.com & wikipedia.org